中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-13

论文引用格式:LIN Xiaoyong, JIAO Kaiyi, WANG Lingfeng. Lightweight mixture-of-experts feed-forward 3D Gaussian splatting model driven by semantic anchor routing[J/OL]. Journal of Image and Graphics, XXXX:1-13. DOI: 10.11834/jig.260121. (林晓勇, 焦恺谊, 汪凌峰. 语义锚点路由驱动的轻量级混合专家前馈式三维高斯泼溅模型[J/OL]. 中国图象图形学报, XXXX:1-13. DOI: 10.11834/jig.260121.)[DOI:10.11834/jig.260121]

语义锚点路由驱动的轻量级混合专家前馈式三维高斯泼溅模型

林晓勇,焦恺谊,汪凌峰

北京化工大学　信息科学与技术学院,北京　100029

摘　要:**目的**　现有前馈式三维高斯泼溅(feed-forward three-dimensional Gaussian splatting,FF-3DGS)模型在稀疏视角输入及外部条件变化下,常面临几何重建不稳定及纹理细节退化等难题。同时,部分方法通过复杂跨视角注意力或特征细化网络增强表征,增加了计算开销;多类高斯属性的联合回归还可能引入任务间的特征竞争。为突破上述性能瓶颈,构建了一种兼顾重建质量与参数规模的轻量级前馈重建框架。**方法**　提出一种语义锚点驱动的块级混合专家前馈模型(semantic-anchor guided mixture-of-experts for feed-forward 3D Gaussian splatting,SAGE-Splat)。该模型首先将单目深度、局部纹理与全局语义等多模态先验对齐至统一表征空间;随后,提取对比语言—图像预训练(contrastive language-image pre-training,CLIP)语义特征作为跨视角稳定锚点,在统一空间内建立图像块级参照系,以促进跨视角下同一语义实体特征分配的一致性;最后,引入任务条件化单层混合专家(mixture of experts,MoE)路由器,通过 Gumbel-TopK 策略稀疏激活特定专家网络,实现对高斯椭球均值、协方差、颜色及不透明度等四类属性的解耦回归。**结果**　实验结果表明,SAGE-Splat 以 15 M 参数在多个数据集上取得了具有竞争力的性能。在 RealEstate10K 数据集上结构相似性达到 0.729,在 DL3DV 数据集上峰值信噪比达到 23.17。**结论**　该模型验证了语义锚点路由与任务条件化专家分配在参数规模和重建质量之间的平衡作用,为高质量稀疏视角三维重建提供了一种轻量化实现路径。

关键词:三维重建;高斯泼溅;多任务学习;混合专家模型;轻量化架构

Lightweight mixture-of-experts feed-forward 3D Gaussian splatting model driven by semantic anchor routing

LIN Xiaoyong, JIAO Kaiyi, WANG Lingfeng

College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract: **Objective** Feed-forward three-dimensional Gaussian splatting (FF-3DGS) predicts Gaussian primitives directly from sparse input views in a single forward pass, avoiding the per-scene optimization required by conventional 3D Gaussian splatting and neural radiance fields (NeRF). Existing FF-3DGS models still suffer from three problems. Under sparse views and changing external conditions, the same semantic entity is often encoded inconsistently across views, which leads to unstable geometry and degraded texture details. Some approaches fuse depth, texture, and semantic features with heavy

收稿日期:2026-03-06;**修回日期**:2026-10-29

*通信作者:汪凌峰,男,教授,博士生导师,研究方向为机器学习理论和计算机视觉。Email:2021500101@buct.edu.cn

基金项目:中国国家自然科学基金(项目编号:62376017);北京市自然科学基金资助项目(项目编号:L253019)

Supported by: National Natural Science Foundation of China (Grant No. 62376017); Beijing Natural Science Foundation (Grant No. L253019)

© 中国图象图形学报版权所有

cross-view attention or refinement networks, introducing large computational overhead. A shared representation is usually used to regress the Gaussian mean, covariance, color, and opacity jointly, so the four attribute-prediction tasks interfere with each other. This study examines semantic features as routing anchors in a lightweight framework that balances reconstruction quality and model size. **Methods** We propose a semantic-anchor guided mixture-of-experts framework for feed-forward 3D Gaussian splatting (SAGE-Splat). First, multimodal priors, namely monocular depth, local texture, and global semantics, are aligned into a unified representation space through channel embedding and spatial resizing, so that the three branches share the same feature grid with 128 channels. Second, semantic features from the contrastive language-image pre-training (CLIP) model are extracted as stable anchors across views. Instead of concatenating CLIP features as an additional input, we project CLIP patch tokens through a linear layer and layer normalization to form routing queries, and build a patch-level reference frame in the unified space, so that patches belonging to the same semantic entity in different views tend to follow consistent expert-assignment paths. Third, a task-conditioned single-layer mixture of experts (MoE) router is introduced. For each patch, four task gates corresponding to opacity, position, scale-related expansion parameters, and color map the shared semantic query to expert scores. Gumbel-TopK selects two experts per task. A straight-through estimator uses hard weights in the forward pass and a Softmax relaxation with a temperature of 0.8 in backpropagation, allowing task-specific expert combinations within a shared expert pool. Each expert processes the depth, semantic, and texture branches and concatenates their outputs along the channel dimension. Task-specific prediction heads then recover the spatial resolution and output one channel for opacity, three for position, thirteen for the intermediate expansion representation, and three for color. The expansion representation is mapped to a lower-triangular factor so that the resulting covariance matrix is strictly positive definite. Gaussians predicted from all views are transformed into a common world coordinate system using the camera extrinsics, projected onto target views, and composited by differentiable rasterization. Training uses a weighted combination of mean squared error (MSE) and learned perceptual image patch similarity (LPIPS) losses. Experiments are conducted on the RealEstate10K and DL3DV benchmarks with four reference views. All baselines are retrained under identical conditions on a single NVIDIA RTX 3090 graphics card for 30 000 iterations with a total batch size of 4, using the AdamW optimizer, a cosine learning-rate schedule, and an initial learning rate of 0.001. The input resolution is 256 pixels by 256 pixels for RealEstate10K and 256 pixels by 448 pixels for DL3DV. For DL3DV, 1 000 iterations of RealEstate10K pretraining precede further training with the backbone frozen. **Results** With only 15 million parameters, SAGE-Splat achieves competitive reconstruction quality on both benchmarks. On RealEstate10K it reaches a structural similarity (SSIM) of 0.729, a peak signal-to-noise ratio (PSNR) of 22.23 dB, and a learned perceptual image patch similarity of 0.228, outperforming DepthSplat, MvSplat, PixelSplat, and OmniSplat as well as several neural radiance field baselines, most of which have far more parameters; DepthSplat, for example, has 128 million. On DL3DV it reaches a structural similarity of 0.714, a peak signal-to-noise ratio of 23.17 dB, and a perceptual score of 0.312, whereas DepthSplat and PixelSplat run out of memory under the same hardware constraint and MvSplat scores clearly lower. Qualitative comparisons on RealEstate10K show more stable edges and textures and lower pixel-wise absolute error than MvSplat and DepthSplat, especially near high-contrast boundaries. Ablation studies confirm that every component contributes: removing the CLIP router causes the largest degradation, with structural similarity falling to 0.371, and removing the mixture-of-experts module, the depth branch, the global semantic branch, or the texture branch also worsens reconstruction quality across the three metrics. Visualization with t-distributed stochastic neighbor embedding shows that the four tasks' features are initially interleaved and progressively form distinct cluster cores after 10 000 and 30 000 iterations. This qualitative pattern is consistent with more task-specific feature assignment. **Conclusions** The experiments indicate that semantic-anchor routing and task-conditioned expert assignment can balance parameter scale and reconstruction quality within the evaluated four-view protocol. The model provides a lightweight implementation path for high-quality three-dimensional reconstruction under limited computing resources. Its performance at very high resolutions, with large numbers of input views, and under large cross-domain shifts remains to be verified, and future work will explore self-supervised large-scale pretraining and memory-efficient training for high-resolution scenes.

Key words: 3D reconstruction; Gaussian splatting; multi-task learning; mixture of experts; lightweight architecture

© 中国图象图形学报版权所有

Lin Xiaoyong, Jiao Kaiyi, Wang Lingfeng. Lightweight mixture-of-experts feed-forward 3D Gaussian splatting model driven by semantic anchor routing [J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260121. (林晓勇,焦恺谊,汪凌峰. 语义锚点路由驱动的轻量级混合专家前馈式三维高斯泼溅模型 [J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260121.)

0 引 言

高保真、实时的三维重建是连接物理世界与数字孪生的关键环节,在虚拟现实、自动驾驶、机器人与工业巡检等场景中具有重要价值。神经辐射场(neural radiance fields,NeRF)通过隐式表征实现了高质量新视图合成,但其训练与推理代价高,难以满足实时交互需求(Mildenhall 等,2022;Chen T 等,2025)。三维高斯泼溅(three-dimensional Gaussian splatting,3DGS)以显式高斯椭球表征结合可微光栅化渲染,实现了高质量与实时性的兼顾(Kerbl 等,2023)。有关3DGS的技术进展、应用与挑战,可参见相关综述(Wu 等,2024;Bao 等,2025;Fei 等,2025)。国内相关研究也在持续推进:曾志鸿等(2025)利用本征属性分类改善神经辐射场的视角及语义一致性;吕恒烨等(2025)研究三维高斯泼溅框架下的室内场景多视图本征分解。这些工作为跨视角一致表征及几何—外观解耦提供了相关参考。

目前较优的3DGS模型依赖单场景优化范式,围绕单场景优化这一主题,研究者主要从更可靠的几何初始化、约束与更强的观测一致性损失两条路径改进。借鉴由深度信息引导的稀疏视角 NeRF 模型的经验(Tan 等,2024;Wang 等,2024;Wang 等,2023),一系列工作通过引入单目或多目深度估计先验来约束高斯中心初始化与剪枝致密化,例如 DNGaussian (Li 等, 2024) 与 Depth-Aware 3DGS (Kumar 和 Vats,2024)。此外,研究者还探索了观测损失与场景致密化策略(Jiang 等,2025;Peng 等,2024),以及借助光学变换生成补充视点并设计视图匹配损失以提升稀疏视角几何一致性(Han 等,2024)。尽管这些方法在特定场景下能显著提升重建精度,但通常需要面向场景的损失设计与多阶段优化,迁移到新域时仍存在不稳定与调参成本高的问题。

为突破单场景优化泛化性不足的天然瓶颈,前馈式三维高斯泼溅(feed-forward three-dimensional Gaussian splatting,FF-3DGS)尝试从稀疏视角图像直接回归高斯参数,实现一次前向推理的即时重建。例如,PixelSplat(Charatan 等,2024)通过沿相机射线预测离散概率分布并将概率分布作为特征来缓解椭球定位的局部最优;MvSplat(Chen 等,2024)引入平面扫描代价体以显式构建跨视角几何线索。进一步地,DepthSplat(Xu H 等,2025)结合 Depth Anything(Yang 等,2024)将深度估计特征与高斯泼溅耦合以提升几何鲁棒性,NexusGS(Zheng 等,2025)则通过对极匹配与光流增强深度特征的尺度鲁棒性以此提高模型的重建性能。FreeSplatter(Xu J 等,2025)则采用前馈框架,从未标定位姿的稀疏视图预测高斯表示并估计相机参数。尽管上述工作通过为参数回归网络引入更多特征提高了模型性能,现有 FF-3DGS 在稀疏视角与条件变化下仍常出现重建不稳定与细节退化。

区别于以上工作将退化主要归因于稀疏视角导致的“特征不足”,机制分析表明,现有模型缺乏跨视角一致的特征分配机制,即同一语义实体在不同视角下的局部表征不一致,进而回归出不稳定的高斯参数。除此之外,多源模态对齐代价高且参数预测过程耦合问题也限制了模型的泛化能力,具体表现为现有前馈模型为融合深度、纹理与语义往往引入重型 refinement 网络或跨视角 Transformer,显著增加计算开销,并在实践中更依赖对预训练模块的再训练。

基于上述观察,本研究将研究重点从“继续增加特征通道”转向“先建立跨视角稳定参照系,再据此进行特征分配与参数回归”。与直接将对比语言—图像预训练(contrastive language-image pre-training,CLIP)特征作为附加输入的做法不同,本文将 CLIP 语义 token 作为图像块级路由锚点,以利用视觉语言预训练获得的语义表征(Radford 等,2021;Pan 等,2022),并以此促进跨视角专家分配的一致性。在此基础上提出语义锚点驱动的块级混合专家前馈高斯泼溅框架——semantic-anchor guided mixture-of-experts for feed-forward 3D Gaussian splatting(SAGE-Splat)。该框架并非简单叠加 CLIP 与混合专家(mixture of experts,MoE),而是将语义锚点用于跨视

© 中国图象图形学报版权所有

角路由约束，并通过任务条件化单层 MoE 为均值 $\boldsymbol{\mu}$、协方差 $\boldsymbol{\Sigma}$、颜色 $\boldsymbol{c}$ 与不透明度 α 分配差异化专家组合，从而同时缓解跨视角表征漂移与高斯参数任务耦合。

论文主要贡献如下：

1）提出一种面向 FF-3DGS 的语义锚点驱动块级路由策略。不同于将 CLIP 特征直接作为额外输入或在特征层面进行简单加权融合，本文将 CLIP patch token 用作专家路由查询，使语义相近图像块在跨视角条件下更倾向于进入一致的专家计算路径，从而缓解同一语义实体跨视角特征分配不一致的问题。

2）提出任务条件化的单层 MoE 高斯参数预测框架。该框架在块级粒度上为均值、协方差、颜色和不透明度分配差异化专家组合，以降低共享特征直接回归多类高斯参数时的任务耦合。

3）在受限计算条件下验证该轻量化设计的有效性。实验表明，在本文实验设置下，该方法以较低参数量取得了具有竞争力的新视图合成质量。

1 相关工作

1.1 稀疏视角三维高斯泼溅

稀疏视角三维重建通常面临观测不足、遮挡区域歧义和几何初始化不稳定等问题。NeRF 系列方法通过隐式辐射场实现高质量新视图合成，但训练和推理开销较高；3DGS 以显式高斯椭球结合可微光栅化提升了渲染效率。近期方法通过深度先验、结构一致性约束、匹配先验或视图一致性损失改善稀疏视角下的稳定性，但多数仍依赖单场景优化或多阶段调参，泛化到新场景时存在额外成本。

1.2 前馈式 3DGS 与跨视角特征聚合

前馈式 3DGS 试图从稀疏输入图像一次前向预测高斯参数。PixelSplat 通过射线方向概率分布缓解高斯中心定位问题，MvSplat 利用平面扫描代价体显式构建跨视角几何线索，DepthSplat 将深度估计特征与高斯预测结合，NexusGS 进一步利用对极深度先验增强稀疏视角重建。上述方法主要围绕几何线索和跨视角匹配增强展开，但在视角变化、遮挡和外观扰动下，同一语义实体仍可能被分配到不一致的局部表征路径。Hi-Gaussian（Xie 等，2025）面向单视图前馈式重建，通过归一化球面投影拓展视域，并以“领航—跟随”层次化高斯采样补全遮挡区域。DGOcc（Zhao 等，2025）利用显式深度上下文、全局查询和分层选择性监督，兼顾单目三维占用预测的精度与效率。二者在视域拓展、遮挡补全和深度先验组织方面的探索，为语义锚点与几何先验协同及遮挡感知路由提供了参考。

1.3 语义表征与混合专家模型

CLIP 提供视觉语言对齐表征（Radford 等，2021），DINOv2 等自监督模型则提供通用视觉特征（Oquab 等，2023）。MoE 结构用于动态容量分配与任务关系建模（Ma 等，2018；Zhou Y 等，2022）；多任务学习中的共享表征与梯度干扰问题也受到持续关注（Yu J 等，2025；Yu T 等，2020）。已有语义 3DGS 方法通常将 CLIP 或语言语义特征用于语义理解、开放词汇分割、语义场构建或跨视角语义监督；已有 MoE 方法则多将专家网络作为通用任务解耦模块。本文关注的不是语义场渲染或开放词汇识别，而是将 CLIP patch token 放入 FF-3DGS 高斯参数预测前的路由决策环节，使语义相近的跨视角图像块更倾向于共享一致专家选择，同时利用任务条件化 MoE 为不同高斯参数预测任务分配不同专家组合，避免简单特征拼接或加权融合造成的任务竞争。MR-COSMO（Li C 等，2026）通过直接跨模态对齐和视觉—文本记忆召回，显式关联点云、图像与文本以服务查询驱动三维分割。尽管任务不同，该工作提示语义表征除作为附加特征外，还可参与按需信息调用，为进一步厘清本文的语义锚点路由及未来构建可记忆的语义—几何路由提供了参考。

1.4 与相近工作的区别

CLIP-GS 等工作主要利用 CLIP 语义信息提升 3DGS 的语义理解和视图一致性（Liao 等，2024），SemanticSplat 则面向几何、外观与语义场的联合前馈重建，并支持开放词汇或可提示分割（Li Q 等，2025）。本文不以语义分割或开放词汇场景理解为主要目标，而是将语义特征用于高斯参数回归前的路由决策，服务于稀疏视角 FF-3DGS 中的跨视角特征分配一致性与参数预测解耦。

Minute4D（Xie 等，2026）面向动态四维高斯泼溅（four-dimensional Gaussian splatting，4DGS）的分钟级高保真训练，为后续探索时序一致的专家分配提供了参考。总体上，Hi-Gaussian、DGOcc、MR-COSMO 和 Minute4D 分别推进了单视图遮挡补全、

© 中国图象图形学报版权所有

单目占用推断、跨模态对齐和动态建模；这些工作共同表明有限观测三维建模需要同时关注几何补全、语义组织和计算资源分配，并为 SAGE-Splat 向单视图补全与动态场景扩展提供了后续启示。本文则面向静态四视图前馈重建，以 CLIP 语义锚点稳定跨视角块级路由，并以任务条件化 MoE 解耦四类高斯参数预测，因而在输入协议、任务目标和核心机制上与其互补。

为进一步厘清相近工作的技术差异，本文围绕输入与场景、建模范式、任务目标、核心机制和评价可比性5个方面梳理三者的研究侧重。输入与场景方面，Hi-Gaussian 面向单张图像的静态三维重建，SAGE-Splat 采用4个参考视图进行静态稀疏视图重建，Minute4D 关注 4D 高斯建模；建模范式方面，Hi-Gaussian 与 SAGE-Splat 均采用前馈预测，Minute4D 强调高保真 4D 高斯表示的快速训练；任务目标方面，Hi-Gaussian 重点处理输入视域外及遮挡区域，SAGE-Splat 关注静态稀疏视图的新视图合成，Minute4D 则将高保真建模与训练效率作为研究重点；核心机制方面，Hi-Gaussian 利用归一化球面投影与层次化高斯采样，SAGE-Splat 采用 CLIP 语义锚点路由与任务条件化 MoE。评价可比性方面，四视图静态重建、单图补全与 4D 高斯建模并非同一评价任务，直接并列各论文原始数值不能构成受控的等条件比较。因此，本文通过研究目标及协议条件的比较说明三者的联系与差异，不根据跨协议数值作性能排序。

2 方　法

SAGE-Splat 由多源表征对齐、语义锚点路由和高斯参数预测组成，总体结构如图 1 所示。每个参考视图的图像块分别为不透明度、位置、扩展参数和颜色四个任务选择专家。四个任务共享同一专家库，通过任务条件化混合专家（task-conditioned mixture of experts，Task-MoE）路由形成不同的专家组合；CLIP 语义查询用于引导专家选择，促进语义相关图像块的特征分配一致性。

下文对任一参考视图说明运算，并省略视图下标。各视图共享网络参数，预测高斯经相机外参变换到公共世界坐标后组合。

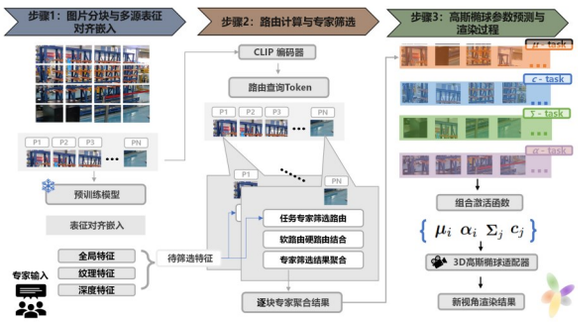


图 1　SAGE-Splat 基本结构图

Figure 1　SAGE-Splat basic structure diagram

2.1　模态特征对齐与逐块语义锚点构造

深度、全局语义和局部纹理三类输入特征分别记为 $\boldsymbol{F}$ 的 d、s、t 分支。各分支先进行 1 × 1 卷积通道变换、多层感知机（multilayer perceptron，MLP）嵌入和空间尺寸调整，以保持一致的特征网格与通道数。

将通道嵌入和空间尺寸调整合记为模态映射 $\boldsymbol{\Phi}$，得到式（1）。m 分别表示深度、语义和纹理模态，H_{f} 与 W_{f} 为统一特征图的高宽，通道数 C 设为 128。

$$\boldsymbol{X}^{m}=\boldsymbol{\Phi}_{m}\left(\boldsymbol{F}^{m}\right)\in\mathbf{R}^{H_{\mathrm{f}}\times W_{\mathrm{f}}\times C} \tag{1}$$

式中，$\boldsymbol{\Phi}$ 使三类输入特征具有一致的通道数和空间网格。令 p 为图块边长，特征图高宽均为 p 的整数倍，以便进行完整的图块划分与恢复。

每个模态被划分为 $N_{\mathrm{p}}=\dfrac{H_{\mathrm{f}}W_{\mathrm{f}}}{p^{2}}$ 个图块，i 表示图块索引。对应图块在三分支中使用相同的空间范围；CLIP token 也对齐到该图块网格。

2.2　任务条件化特征选择路由器

对于图块 i，令 $\boldsymbol{z}$ 为对应的 CLIP 语义 token，经过线性投影和层归一化（layer normalization，LN）得到查询 $\boldsymbol{q}$。任务标识 k 分别取 α、μ、ε、c，代表不透明度、位置、扩展参数和颜色；任务 k 的线性门控 $\boldsymbol{G}$ 将查询映射到 E 个专家分数。

共享语义表征为路由提供统一的查询空间。任务标识 k 用于区分四组参数预测门控，每个图块均生成全部四类高斯属性所需的表征。

与仅将 CLIP 作为额外特征通道的方式相比，本方法将语义查询用于专家分数计算。各参数任务可以选择不同专家组合，在共享专家容量的同时实现差异化特征聚合。

$$\begin{aligned}\boldsymbol{q}_{i}&=\mathrm{LN}\left(\boldsymbol{W}_{\mathrm{q}}\boldsymbol{z}_{i}\right)\\ \boldsymbol{r}_{i}^{k}&=\boldsymbol{G}_{k}\left(\boldsymbol{q}_{i}\right)\in\mathbf{R}^{E}\end{aligned} \tag{2}$$

© 中国图象图形学报版权所有

式中，$\boldsymbol{W}_{\mathrm{q}}$为查询投影，LN为层归一化；各任务门控输出维数均为专家总数E。其中，$\boldsymbol{r}_i^k$为任务k的专家分数向量，$r_{i,j}^k$为第j个分量，$\tilde{r}_{i,j}^k$为加入Gumbel扰动后的分数。

$$\begin{aligned}\tilde{r}_{ij}^k &= r_{ij}^k + g_{ij}^k \\ g_{ij}^k &= -\log\left(-\log\left(u_{ij}^k\right)\right)\end{aligned} \tag{3}$$

式中，u在(0, 1)上独立均匀采样，g为Gumbel扰动。训练阶段使用式(3)的扰动分数；推理阶段令g为0，采用确定性选择。后续Softmax连续松弛的温度τ设为0.8。

每个任务独立选取分数最高的K个专家，形成式(4)中的索引集合。K设为2，专家总数E满足$E \geqslant K$；并列分数采用固定专家索引顺序处理。

$$\boldsymbol{S}_i^k = \mathrm{TopK}\left(\tilde{\boldsymbol{r}}_i^k, K\right), K = 2 \tag{4}$$

式中，集合中的元素表示专家索引。四个参数任务分别产生大小为K的专家集合，并利用所选专家输出构建各自的任务特征。

Gumbel扰动后的Top-K用于离散专家选择(Kool等，2019)。选中专家的硬权重定义为1/K，Softmax提供温度控制的连续松弛(Jang等，2017)，二者分别见式(5)—(6)。

$$h_{ij}^k = \frac{1\left\{j \in \boldsymbol{S}_i^k\right\}}{K} \tag{5}$$

$$s_{ij}^k = \frac{\exp\left(\dfrac{\tilde{r}_{ij}^k}{\tau}\right)}{\sum_{l=1}^{E}\exp\left(\dfrac{\tilde{r}_{i,l}^k}{\tau}\right)} \tag{6}$$

式中，指示函数在专家被选中时为1，否则为0；式(6)的归一化覆盖全部E个候选专家。硬权重与软权重对应同一路由的前向选择和连续代理。

为处理离散选择的不可微性，采用式(7)的直通估计。sg表示停止梯度操作：前向保持输入值，反向截断该项梯度。

$$w_{ij}^k = \mathrm{sg}\left(h_{ij}^k - s_{ij}^k\right) + s_{ij}^k \tag{7}$$

式中，前向计算采用硬权重，每个任务恰有K个非零权重且权重和为1；反向通过Softmax软权重传递代理梯度，使门控分数能够参与联合优化。

2.3 多任务高斯参数预测模块

每个专家j包含深度、语义和纹理三个分支，式(8)—(10)分别定义其图块输出。三个输出具有相同空间尺寸，并按式(11)沿通道维拼接，构成该专家的多模态表征。其中，$\boldsymbol{P}_i^m$为模态m的第i个输入图块，E_j^m为第j个专家的该模态分支，$\boldsymbol{U}_{i,j}^m$为其输出特征。

$$\boldsymbol{U}_{ij}^{\mathrm{d}} = E_j^{\mathrm{d}}\left(\boldsymbol{P}_i^{\mathrm{d}}\right) \tag{8}$$

$$\boldsymbol{U}_{ij}^{\mathrm{s}} = E_j^{\mathrm{s}}\left(\boldsymbol{P}_i^{\mathrm{s}}\right) \tag{9}$$

$$\boldsymbol{U}_{ij}^{\mathrm{t}} = E_j^{\mathrm{t}}\left(\boldsymbol{P}_i^{\mathrm{t}}\right) \tag{10}$$

$$\boldsymbol{U}_{ij} = \mathrm{Concat}\left(\boldsymbol{U}_{ij}^{\mathrm{d}}, \boldsymbol{U}_{ij}^{\mathrm{s}}, \boldsymbol{U}_{ij}^{\mathrm{t}}\right) \tag{11}$$

$$\boldsymbol{F}^k = \mathrm{Unpatch}\left(\left\{\boldsymbol{\Theta}_i^k\right\}_{i=1}^{N_{\mathrm{p}}}\right) \tag{12}$$

式中，$\boldsymbol{\Theta}$表示任务加权后的图块特征，四个任务的展开分别见式(13)—(16)。Unpatch将全部图块按原网格恢复为空间特征图。四个任务使用同一专家库，并通过各自的路由权重形成不同的特征组合。

$$\boldsymbol{\Theta}_i^{\alpha} = \sum_{j=1}^{E} w_{ij}^{\alpha}\boldsymbol{U}_{ij} \tag{13}$$

$$\boldsymbol{\Theta}_i^{\mu} = \sum_{j=1}^{E} w_{ij}^{\mu}\boldsymbol{U}_{ij} \tag{14}$$

$$\boldsymbol{\Theta}_i^{\varepsilon} = \sum_{j=1}^{E} w_{ij}^{\varepsilon}\boldsymbol{U}_{ij} \tag{15}$$

$$\boldsymbol{\Theta}_i^{\mathrm{c}} = \sum_{j=1}^{E} w_{ij}^{\mathrm{c}}\boldsymbol{U}_{ij} \tag{16}$$

各任务采用三层卷积预测头D，并恢复目标空间分辨率。对每个空间位置u分别预测高斯属性，不透明度、位置、扩展参数和颜色的输出通道数依次为1、3、13、3。各预测头保留空间维度，逐位置映射见式(17)。

$$\begin{aligned}\alpha_u &= \mathrm{sigmoid}\left(D_{\alpha}\left(\boldsymbol{F}^{\alpha}\right)_u\right) \\ \boldsymbol{\mu}_u^{\mathrm{cam}} &= \boldsymbol{b}_{\mu} + \boldsymbol{S}_{\mu}\tanh\left(D_{\mu}\left(\boldsymbol{F}^{\mu}\right)_u\right) \\ \boldsymbol{\varepsilon}_u &= \mathrm{ReLU}\left(D_{\varepsilon}\left(\boldsymbol{F}^{\varepsilon}\right)_u\right) \\ \boldsymbol{c}_u &= \frac{1 + \tanh\left(D_{\mathrm{c}}\left(\boldsymbol{F}^{\mathrm{c}}\right)_u\right)}{2}\end{aligned} \tag{17}$$

式中，Sigmoid将不透明度约束在[0, 1]，颜色经Tanh后平移缩放至[0, 1]。位置通过偏移$\boldsymbol{b}_{\mu}$与正对角尺度$\boldsymbol{S}_{\mu}$进行反归一化，得到源相机坐标中的三维中心。13维$\boldsymbol{\varepsilon}$为中间扩展表征，适配映射A_{Σ}将其转换为6个因子参数，并构造式(17a)中的下三角矩阵$\boldsymbol{L}$。

$$\begin{aligned}\boldsymbol{a}_u &= A_{\Sigma}\left(\boldsymbol{\varepsilon}_u\right) \in \mathbf{R}^6 \\ \boldsymbol{L}_u &= \begin{bmatrix} a_1 & 0 & 0 \\ a_2 & a_3 & 0 \\ a_4 & a_5 & a_6 \end{bmatrix}\end{aligned} \tag{17a}$$

$$\boldsymbol{\Sigma}_u^{\mathrm{cam}} = \boldsymbol{L}_u\boldsymbol{L}_u^{\mathrm{T}} + \delta\boldsymbol{I}_3, \delta > 0 \tag{17b}$$

式中，$\boldsymbol{a}_u$为6维因子参数向量，a_1至a_6为其分量；式(17b)中的$\boldsymbol{I}_3$为3阶单位矩阵。式(17b)中δ为严格

© 中国图象图形学报版权所有

正数。对任意非零向量 $\boldsymbol{x}$，有 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x} = \|\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}\|^2 + \delta\|\boldsymbol{x}\|^2 > 0$，因此协方差严格正定，避免了退化因子导致的奇异矩阵。

令源视图从世界坐标到相机坐标的旋转矩阵和平移向量分别为 $\boldsymbol{R}_v$、$\boldsymbol{t}_v$，各视图高斯按式(17c)变换到世界坐标后组合，使中心和协方差处于共同坐标系。

$$\begin{aligned}\boldsymbol{\mu}^{\mathrm{w}} &= \boldsymbol{R}_v^{\mathrm{T}}\left(\boldsymbol{\mu}^{\mathrm{cam}} - \boldsymbol{t}_v\right)\\ \boldsymbol{\Sigma}^{\mathrm{w}} &= \boldsymbol{R}_v^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{cam}}\boldsymbol{R}_v\end{aligned} \tag{17c}$$

式中，上标所标示的两类坐标分别为世界坐标与源相机坐标；旋转矩阵和平移向量沿用源视图的外参定义。

2.4 渲染过程与损失函数

对目标相机，$\boldsymbol{R}_{\mathrm{t}}$、$\boldsymbol{t}_{\mathrm{t}}$为世界到相机的旋转和平移，$\boldsymbol{K}_{\mathrm{cam}}$为内参矩阵。式(18)先得到相机点 $\boldsymbol{y}$ 和齐次投影，再执行透视除法得到二维中心 $\boldsymbol{m}$；仅对目标相机前方的合法投影点进行渲染。其中，$\tilde{\boldsymbol{p}}$ 为齐次投影坐标。

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{R}_{\mathrm{t}}\boldsymbol{\mu}^{\mathrm{w}} + \boldsymbol{t}_{\mathrm{t}}\\ \tilde{\boldsymbol{p}} &= \boldsymbol{K}_{\mathrm{cam}}\boldsymbol{y}\\ \boldsymbol{m} &= \left(\frac{\tilde{p}_1}{\tilde{p}_3}, \frac{\tilde{p}_2}{\tilde{p}_3}\right)^{\mathrm{T}}\end{aligned} \tag{18}$$

$$\boldsymbol{\Sigma}^{\mathrm{2D}} = \boldsymbol{J}\boldsymbol{R}_{\mathrm{t}}\boldsymbol{\Sigma}^{\mathrm{w}}\boldsymbol{R}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}} \tag{19}$$

式中，$\boldsymbol{J}$ 为目标相机透视映射在 $\boldsymbol{y}$ 处的 2×3 雅可比。式(19)通过局部一阶线性化得到投影协方差。高斯按相机深度排序，n 表示排序后的高斯索引，以区别于专家索引 j。

对像素 $\boldsymbol{p}$，令 $\Delta\boldsymbol{p}$ 表示其与第 n 个投影中心的偏移。像素相关的有效不透明度由不透明度 α 与二维高斯核相乘得到，再结合累积透射率进行颜色合成，如式(20)所示。

$$a_n(\boldsymbol{p}) = \alpha_n \exp\left(-\frac{1}{2}\Delta\boldsymbol{p}_n{}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_n^{\mathrm{2D}}\right)^{-1}\Delta\boldsymbol{p}_n\right) \tag{20}$$

$$\boldsymbol{C}(\boldsymbol{p}) = \sum_{n=1}^{N_{\mathrm{g}}} T_n(\boldsymbol{p})a_n(\boldsymbol{p})\boldsymbol{c}_n$$

$$T_n(\boldsymbol{p}) = \prod_{l=1}^{n-1}\left(1 - a_l(\boldsymbol{p})\right) \tag{21}$$

式中，N_{g} 为参与当前像素合成的高斯数，T 为累积透射率。式(20)给出前景合成项；非黑背景可通过剩余透射率加权加入。前景合成权重之和为1减去最终透射率，不超过1。其中，$a_n(\boldsymbol{p})$ 表示第 n 个高斯在像素 $\boldsymbol{p}$ 处的有效不透明度，$\boldsymbol{C}(\boldsymbol{p})$ 表示合成颜色。

采用均方误差(mean squared error，MSE)与学习感知图像块相似度(learned perceptual image patch similarity，LPIPS)的加权组合作为图像重建损失。式(22)中，各目标视图的两项误差分别记为 ℓ_{MSE} 与 ℓ_{LPIPS}，λ_{p} 为感知损失权重。

$$\begin{aligned}L_{\mathrm{gs}} &= \sum_{m=1}^{M}\ell_m\\ \ell_m &= \ell_{\mathrm{MSE}}^m + \lambda_{\mathrm{p}}\ell_{\mathrm{LPIPS}}^m\end{aligned} \tag{22}$$

式中，总损失记为 L_{gs}，第 m 个目标视图的损失记为 ℓ_m。M 为目标视图数量。MSE 与 LPIPS 均在对应目标视图的预测图像和真实图像之间计算，具体定义见第3.1节。

3 实　验

3.1 评价指标

本文选取结构相似性(structural similarity，SSIM)、峰值信噪比(peak signal-to-noise ratio，PSNR)和学习感知图像块相似度(LPIPS)作为评价指标，以对模型性能进行综合评估。这3个指标分别从局部结构保真度、全局像素级数值精度及感知视觉质量3个维度互为补充，共同用于评价模型的新视图合成质量，提供了多维度的定量分析。

SSIM 从亮度、对比度和结构3个方面衡量生成图像与真实图像的相似度(Wang Z 等，2004)。标准 SSIM 在相同图像上取得最大值1，其结构项允许负值；数值越大表明图像越相似。SSIM 的计算过程如式(23)—(24)所示：

$$\begin{cases} l(\boldsymbol{x},\boldsymbol{y}) = \dfrac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\\ c(\boldsymbol{x},\boldsymbol{y}) = \dfrac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\\ s(\boldsymbol{x},\boldsymbol{y}) = \dfrac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\end{cases} \tag{23}$$

式中，μ_x 和 μ_y 分别表示图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的局部窗口均值；σ_x、σ_y 分别表示图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的局部窗口标准差；σ_{xy} 表示图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的局部窗口协方差；C_1、C_2 和 C_3 均为正的稳定常数。

$$\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y}) = \left[l(\boldsymbol{x},\boldsymbol{y})\right]^{\alpha}\cdot\left[c(\boldsymbol{x},\boldsymbol{y})\right]^{\beta}\cdot\left[s(\boldsymbol{x},\boldsymbol{y})\right]^{\gamma} \tag{24}$$

式中，α,β,γ 为各相似性分量的权重指数；$\left[l(\boldsymbol{x},\boldsymbol{y})\right]$ 表示两幅图像的亮度相似性；$c(\boldsymbol{x},\boldsymbol{y})$ 表示两幅图像的对比度相似性；$s(\boldsymbol{x},\boldsymbol{y})$ 表示两幅图像的结构相似性，数值越大表明局部结构越相似。

© 中国图象图形学报版权所有

PSNR是一种基于均方误差的像素级精度指标，以分贝为单位。PSNR值越高，表明图像质量越好。PSNR的计算如式(25)—(26)所示：

$$\mathrm{MSE}=\frac{1}{mn}\left(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[x(i,j)-y(i,j)\right]^2\right) \tag{25}$$

$$\mathrm{PSNR}=10\cdot\lg\left(\frac{MAX^2}{\mathrm{MSE}}\right) \tag{26}$$

式中，图像的高、宽分别为m和n，i与j为像素坐标索引。MAX表示图像的最大像素值。

LPIPS通过预训练深度网络的多层特征响应衡量图像的感知相似性(Zhang等，2018)，其计算如式(27)—(28)所示：

$$d_l=\frac{1}{H_l W_l}\sum_{h,w}\left\|\boldsymbol{w}_l\odot\left(\hat{\boldsymbol{x}}^l_{hw}-\hat{\boldsymbol{y}}^l_{hw}\right)\right\|_2^2 \tag{27}$$

$$\mathrm{LPIPS}(\boldsymbol{x},\boldsymbol{y})=\sum_l d_l \tag{28}$$

式中，l表示视觉网络模型的层索引；$\hat{\boldsymbol{x}}^l_{hw}$表示推理/渲染图像在第$l$层$(h,w)$位置处的单位归一化特征向量；$\hat{\boldsymbol{y}}^l_{hw}$表示真实图像对应的特征向量；$/\!/\cdot/\!/_2^2$表示$L_2$范数的平方，用于增强对差异的惩罚。其中，$H_l$与$W_l$为第$l$层特征图的高和宽，$\boldsymbol{w}_l$为通道权重向量，$\odot$表示逐元素乘法，$d_l$为该层的感知距离。

3.2 计算约束下的实验设置

本文选用RealEstate10K(Zhou等，2018)和DL3DV(Ling等，2024)作为实验基准，并在单张NVIDIA RTX 3090上重新训练所有基线。预训练骨干网络遵循UniMatch配置(Xu等，2023)。所有实验统一训练30 000次，总批量大小为4，单卡批量大小为1；RealEstate10K分辨率为256×256，DL3DV分辨率为256×448，两个数据集均采用4个参考视图。DL3DV实验先在RealEstate10K上预训练1 000次，随后冻结骨干网络并继续训练30 000次。模型基于PyTorch 2.0实现，采用AdamW优化器和余弦学习率调度策略，初始学习率为0.001。

3.3 有效性对比实验

本文选取的对比基线模型涵盖了当前可复现的最先进的前馈3DGS及NeRF类重建方法，且所有模型均在本地重新进行训练以消除不同硬件设备对实验结果的影响：RealEstate10K和DL3DV数据集上的对比实验结果如表1和表2所示。对比基线中的OmniSplat(Lee等，2025)、MVSNeRF(Chen A等，2021)、CodeNeRF(Jang和Agapito，2021)及PixelNeRF(Yu A等，2021)的结果见表1。

表1 RealEstate10K数据集(4视图)上的实验结果

Table 1 Experimental results on the RealEstate10K dataset (4 views)

类型	模型	参数量/M	SSIM ↑	PSNR ↑	LPIPS ↓
FF-3DGS	**本文**	**15**	**0.729**	**22.23**	**0.228**
	DepthSplat	128	0.689	19.84	0.273
	MvSplat	16	0.653	19.15	0.291
	PixelSplat	114	0.638	18.91	0.305
	OmniSplat	132	0.695	19.78	0.267
NeRF	MVSNeRF	148	0.647	19.19	0.288
	CodeNeRF	108	0.665	19.52	0.279
	PixelNeRF	128	0.589	16.78	0.324

注：加粗数值表示对应列的最优结果；↑、↓分别表示数值越大、越小越好。M表示百万个参数，参数量以较小为优；PSNR的单位为dB。

本文方法在仅15M参数量下取得较优表现，且在感知指标与保真指标上同时领先，体现出较稳定的跨视角重建能力。现有前馈式3DGS基线的主要瓶颈不仅在于“特征强度”不足，也在于缺少稳定的跨视角特征分配依据：PixelSplat、MvSplat等方法强调从像素或代价体中建立几何线索，但在稀疏视角和遮挡条件下仍容易依赖局部纹理匹配，导致同一语义对象在不同视角被编码为不同局部表示；DepthSplat通过深度先验缓解几何不确定性，但深度特征难以单独约束同一语义实体进入一致的参数预测路径。相比之下，SAGE-Splat将语义表征作为路由锚点，在专家选择阶段促使语义相近图像块共享相近的计算路径，从而减少跨视角表征漂移带来的参数不一致；块级稀疏路由为不同图像块及参数预测任务选择相应的专家组合。本文方法在较低参数量下仍能保持具有竞争力的总体性能。

表2给出了DL3DV四视图设置下的结果。SAGE-Splat以15 M参数取得0.714的SSIM、23.17 dB的PSNR和0.312的LPIPS。DepthSplat和PixelSplat在统一硬件条件下发生显存不足(out of memory，OOM)，MvSplat虽可运行但指标较低，说明本文方法在多视图输入下兼顾了重建质量与计算可用性。

为进一步验证SAGE-Splat的有效性，图2展示了RealEstate10K四视图输入设置下，本文方法与

© 中国图象图形学报版权所有

表2　DL3DV数据集(4视图)上的实验结果

Table 2　Experimental results on the DL3DV dataset (4 views)

模型	参数量/M	SSIM ↑	PSNR ↑	LPIPS ↓
本文	15	0.714	23.17	0.312
DepthSplat	128	OOM	OOM	OOM
MvSplat	16	0.502	16.66	0.422
PixelSplat	114	OOM	OOM	OOM

注：加粗数值表示对应列的最优结果；↑、↓分别表示数值越大、越小越好。M表示百万个参数，PSNR的单位为dB。OOM表示在所述硬件条件下显存不足，未获得相应结果。

MvSplat、DepthSplat在同一目标视图上的定性结果。图2(a)给出完整预测图像及局部放大；图2(b)给出共享0—0.25显示范围的红、绿、蓝(red, green, blue，RGB)三通道像素级绝对误差热图。MvSplat在窗框和灯罩边缘存在轻微模糊，DepthSplat在高对比度边界处出现重影与结构扰动；SAGE-Splat的边缘和纹理恢复更稳定，误差响应也更低。

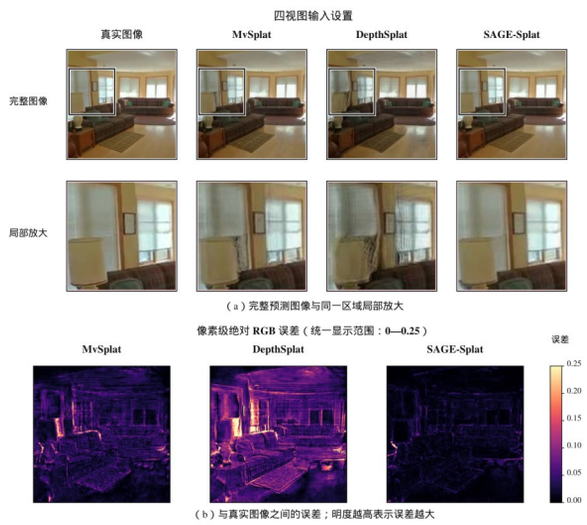


图2　RealEstate10K数据集四视图输入设置下的可视化对比实验：(a)完整预测结果及局部放大；(b)像素级绝对RGB误差热图。

Figure 2　Visualization comparison on RealEstate10K under the 4-view input setting: (a) full predictions and zoomed regions; (b) pixel-wise absolute RGB error maps.

3.4　消融实验与子任务特征独立性分析

为了验证SAGE-Splat模型各核心组件的有效性，本文在RealEstate10K数据集上开展消融实验，结果如表3所示。各对比变体的设置如下：

1) All-Module：包含所有组件的完整模型。

2) w/o-Depth：从多模态输入中移除深度特征分支，并同步剔除其关联的专家处理网络。

3) 视觉Transformer(vision transformer, ViT)消融变体w/o-ViT：从多模态输入中移除全局语义特征分支，并同步剔除其关联的专家处理网络。

4) 卷积神经网络(convolutional neural network, CNN)消融变体w/o-CNN：从多模态输入中移除纹理特征分支，并同步剔除其关联的专家处理网络。

5) w/o-MoE：移除任务选择路由门控网络，直接利用未解耦的完整特征图进行高斯参数预测。

6) w/o-CLIP_Router：不使用CLIP特征作为路由特征，仅使用输入图像自身作为路由计算输入。

表3　消融实验结果

Table 3　Ablation study results

模型	SSIM ↑	PSNR ↑	LPIPS ↓
All-Module	0.729	22.23	0.228
w/o-Depth	0.473	15.05	0.486
w/o-MoE	0.466	14.99	0.479
w/o-ViT	0.467	15.00	0.481
w/o-CLIP_Router	0.371	13.89	0.501
w/o-CNN	0.523	18.53	0.392

注：加粗数值表示对应列的最优结果；↑、↓分别表示数值越大、越小越好；PSNR的单位为dB。All-Module为完整模型，w/o表示移除相应模块。

完整模型(All-Module)在三项指标上取得最优表现，而移除任一组件均导致退化，表明各组件均对整体性能有贡献。去除深度、纹理或全局关系特征后性能均下降，说明多源表征对整体重建质量具有贡献。进一步地，移除MoE后，模型回落为共享表征直接回归多任务参数的范式，难以有效缓解多任务间的特征竞争；当路由不再由CLIP语义锚点驱动时，性能下降最明显，这与语义锚点用于缓解视角、光照和背景变化引起路由波动的设计动机一致。

从替代方案角度看，直接拼接CLIP特征或简单加权融合主要作用于输入表征，未显式改变专家选择的依据；而SAGE-Splat将语义token作为路由查询，使语义信息直接参与Top-K专家选择及后续高斯参数预测。二者的关键区别在于语义信息介入计算路径的位置不同。

© 中国图象图形学报版权所有

表3中的w/o-CLIP_Router与w/o-MoE结果显示，移除任一模块均会造成性能下降，说明CLIP路由与MoE均对整体性能有贡献。结合模型设计，二者分别面向跨视角路由稳定性与多任务参数预测中的特征竞争问题。

为了验证任务条件化特征选择路由器机制的有效性，本文采用t分布随机邻域嵌入（t-distributed stochastic neighbor embedding，t-SNE）结合核密度估计（kernel density estimation，KDE）对特征空间进行了可视化分析。如图3所示，在训练初期，经MoE选择的图像块特征呈现较混杂状态；经过10 000次迭代后，不同任务所选择的表征逐渐形成较清晰的聚类边界；经过30 000次迭代后，对应不同重建子任务的特征在降维空间中形成更明确的聚类簇，其核心分布区域重叠较小。该结果表明，任务条件化路由有助于不同高斯参数预测头选择更匹配的专家表征，从而缓解多任务高斯参数预测中的表征混用问题。

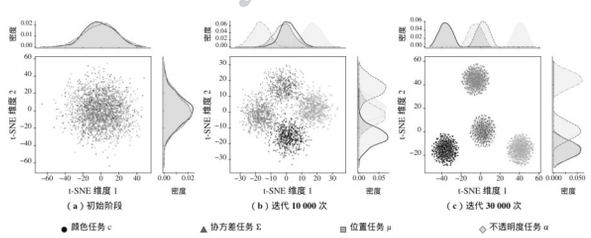


图3　Task-MoE网络中特征选择的有效性分析

Figure 3　Effectiveness analysis of feature selection in the Task-MoE network

4　结　论

针对现有前馈式三维高斯泼溅重建方法在算力受限场景下难以同时兼顾几何精度与渲染质量、且在稀疏视角与域外变化下易出现重建不稳定的问题，本文提出了语义驱动的块级混合专家前馈重建框架SAGE-Splat。该方法以语义表征作为跨视角稳定的锚点，在路由与聚合阶段显式促使同一语义单元共享一致的高斯椭球预测特征，并通过任务条件化的单层MoE路由实现多任务解耦特征分配；在不依赖对骨干网络端到端重训练的前提下，探索了轻量化、高保真、可部署的前馈式3D重建路径。本文主要结论如下：

1）CLIP语义锚点为块级专家选择提供了统一的查询依据。完整模型与去除CLIP路由变体的消融结果表明，语义路由对本文实验设置下的新视图合成质量具有积极作用。

2）任务条件化路由为不同高斯参数预测任务分配差异化专家组合。完整模型与去除MoE变体的比较表明，这一设计有助于提高整体重建质量；特征可视化进一步呈现了不同任务表征的分化。

3）本文模型以15 M参数在RealEstate10K和DL3DV基准上取得了有竞争力的重建结果，说明语义引导的专家分配能够兼顾参数规模与重建质量，为轻量化前馈三维重建提供一种实现路径。

尽管本文方法在参数规模与重建质量之间取得了良好的平衡，但仍受限于当前的实验硬件环境与算法框架。一方面，受训练设备显存容量限制，模型在4K级高分辨率或大规模多视图输入下的极限性能尚未得到充分验证，在极精细纹理的恢复上可能存在瓶颈；另一方面，模型对输入特征的依赖性较强，当前的特征提取仍基于特定数据集预训练的骨干网络，在跨域大差异场景下的泛化鲁棒性仍有提升空间。

未来的研究将着重探索自监督或大规模联合预训练策略，以降低对特定先验的依赖，增强模型的通用泛化能力；同时，将研究更高效的显存优化技术，以支持高分辨率场景下的端到端训练与推理，进一步拓展该方法在工业级大规模场景重建中的应用潜力。

致谢：本研究工作得到中国国家自然科学基金（62376017）及北京市自然科学基金（L253019）的资助。在此对提供资助的机构表示衷心的感谢！

参考文献（References）

Bao Y, Ding T, Huo J, Liu Y, Li Y, Li W, et al. 2025. 3D Gaussian splatting: survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology, 35(7): 6832-6852 [DOI: 10.1109/TCSVT.2025.3538684]

Charatan D, Li S L, Tagliasacchi A and Sitzmann V. 2024. PixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 19457-19467 [DOI: 10.1109/CVPR52733.2024.01840]

Chen A, Xu Z, Zhao F, Zhang X, Xiang F, Yu J, et al. 2021. MVS-

© 中国图象图形学报版权所有

NeRF: fast generalizable radiance field reconstruction from multi-view stereo//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 14104-14113 [DOI: 10.1109/ICCV48922.2021.01386]

Chen T, Yang Q L and Chen Y. 2025. Overview of NeRF technology and applications. Journal of Computer-Aided Design & Computer Graphics, 37(1): 51-74 (陈涛, 杨启亮, 陈寅. 2025. 神经辐射场技术及应用综述. 计算机辅助设计与图形学学报, 37(1): 51-74) [DOI: 10.3724/SP.J.1089.2023-00759]

Chen Y, Xu H, Zheng C, Zhuang B, Pollefeys M, Geiger A, et al. 2024. MVSplat: efficient 3D Gaussian splatting from sparse multi-view images//European Conference on Computer Vision. Cham: Springer: 370-386 [DOI: 10.1007/978-3-031-72664-4_21]

Fei B, Xu J, Zhang R, Zhou Q, Yang W and He Y. 2025. 3D Gaussian splatting as a new era: a survey. IEEE Transactions on Visualization and Computer Graphics, 31(8): 4429-4449 [DOI: 10.1109/TVCG.2024.3397828]

Han L, Zhou J, Liu Y S and Han Z. 2024. Binocular-guided 3D Gaussian splatting with view consistency for sparse view synthesis// Advances in Neural Information Processing Systems, 37. Red Hook: Curran Associates: 68595-68621 [DOI: 10.52202/079017-2191]

Jang E, Gu S and Poole B. 2017. Categorical reparameterization with Gumbel-softmax//Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview. net. https://openreview.net/forum? id=rkE3y85ee

Jang W and Agapito L. 2021. CodeNeRF: disentangled neural radiance fields for object categories//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 12929-12938 [DOI: 10.1109/ICCV48922.2021.01271]

Jiang C, Gao R, Shao K, Wang Y, Xiong R and Zhang Y. 2025. LI-GS: Gaussian splatting with LiDAR incorporated for accurate large-scale reconstruction. IEEE Robotics and Automation Letters, 10(2): 1864-1871 [DOI: 10.1109/LRA.2024.3522846]

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): 139:1-139:14 [DOI: 10.1145/3592433]

Kool W, van Hoof H and Welling M. 2019. Stochastic beams and where to find them: the Gumbel-Top-k trick for sampling sequences without replacement//Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, 97). Long Beach: PMLR: 3499-3508. https://proceedings. mlr. press/v97/kool19a.html

Kumar R and Vats V. 2024. Few-shot novel view synthesis using depth aware 3D Gaussian splatting [EB/OL]. (2024-10-14) [2026-09-05]. https://arxiv. org/abs/2410.11080v1 [DOI: 10.48550/arXiv. 2410.11080]

Lee S, Chung J, Kim K, Huh J, Lee G, Lee M, et al. 2025. OmniSplat: taming feed-forward 3D Gaussian splatting for omnidirectional images with editable capabilities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 16356-16365 [DOI: 10.1109/CVPR52734.2025.01525]

Li C, Zhang P and Wu Y. 2026. MR-COSMO: visual-text memory recall and direct cross-modal alignment method for query-driven 3D segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 40(8): 6010-6018 [DOI: 10.1609/aaai.v40i8.37524]

Li J, Zhang J, Bai X, Zheng J, Ning X, Zhou J, et al. 2024. DNGaussian: optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 20775-20785 [DOI: 10.1109/CVPR52733.2024.01963]

Li Q, Sun J, An L, Su Z, Zhang H and Liu Y. 2025. SemanticSplat: feed-forward 3D scene understanding with language-aware Gaussian fields [EB/OL]. (2025-06-13) [2026-09-05]. https://arxiv. org/abs/2506.09565v2 [DOI: 10.48550/arXiv. 2506.09565]

Liao G, Li J, Bao Z, Ye X, Wang J, Li Q, et al. 2024. CLIP-GS: CLIP-informed Gaussian splatting for real-time and view-consistent 3D semantic understanding [EB/OL]. (2024-04-22) [2026-09-05]. https://arxiv. org/abs/2404.14249v1 [DOI: 10.48550/arXiv. 2404.14249]

Ling L, Sheng Y, Tu Z, Zhao W, Xin C, Wan K, et al. 2024. DL3DV-10K: a large-scale scene dataset for deep learning-based 3D vision//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 22160-22169 [DOI: 10.1109/CVPR52733.2024.02092]

Lyu H Y, Liu Y L, Li H, Yuan X and Xing G Y. 2025. Multi-view intrinsic decomposition of indoor scenes under a 3D Gaussian splatting framework. Journal of Image and Graphics, 30(7): 2514-2527 (吕恒烨,刘艳丽,李宏,袁霞,邢冠宇. 2025. 三维高斯泼溅框架下的室内场景多视图本征分解. 中国图象图形学报, 30(7): 2514-2527) [DOI: 10.11834/jig.240505]

Ma J, Zhao Z, Yi X, Chen J, Hong L and Chi E H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM: 1930-1939 [DOI: 10.1145/3219819.3220007]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2023. DINOv2: learning robust visual features without supervision [EB/OL]. (2023-04-14) [2026-09-05]. https://arxiv. org/abs/2304.07193v1 [DOI: 10.48550/arXiv. 2304.07193]

© 中国图象图形学报版权所有

Pan X, Ye T, Han D, Song S and Huang G. 2022. Contrastive language-image pre-training with knowledge graphs//Advances in Neural Information Processing Systems, 35. Red Hook: Curran Associates: 22895-22910 [DOI: 10.52202/068431-1664]

Peng R, Xu W, Tang L, Liao L, Jiao J and Wang R. 2024. Structure consistent Gaussian splatting with matching prior for few-shot novel view synthesis//Advances in Neural Information Processing Systems, 37. Red Hook: Curran Associates: 97328-97352 [DOI: 10.52202/079017-3087]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, 139). Virtual: PMLR: 8748-8763. https://proceedings.mlr.press/v139/radford21a.html

Tan Z, Zhou Z, Ge Y, Wang Z, Chen X and Hu D. 2024. TD-NeRF: novel truncated depth prior for joint camera pose and neural radiance field optimization//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi: IEEE: 372-379 [DOI: 10.1109/IROS58592.2024.10802634]

Wang B, Zhang D, Su Y and Zhang H. 2024. Enhancing view synthesis with depth-guided neural radiance fields and improved depth completion. Sensors, 24(6): 1919 [DOI: 10.3390/s24061919]

Wang J, Xiao J, Zhang X, Xu X, Jin T and Jin Z. 2023. Depth-based dynamic sampling of neural radiation fields. Electronics, 12(4): 1053 [DOI: 10.3390/electronics12041053]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]

Wu T, Yuan Y J, Zhang L X, Yang J, Cao Y P, Yan L Q, et al. 2024. Recent advances in 3D Gaussian splatting. Computational Visual Media, 10(4): 613-642 [DOI: 10.1007/s41095-024-0436-y]

Xie B, Zhang P, Wei H and Wu Y. 2025. Hi-Gaussian: hierarchical gaussians under normalized spherical projection for single-view 3D reconstruction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE: 28664-28673 [DOI: 10.1109/ICCV51701.2025.02662]

Xie B, Shi C, Zhang P and Wu Y. 2026. Minute4D: training high-fidelity 4D Gaussian splatting in one minute//Computer Vision – ECCV 2026. Cham: Springer Nature Switzerland: 116-133 [DOI: 10.1007/978-3-032-37464-6_7]

Xu H, Peng S, Wang F, Blum H, Barath D, Geiger A, et al. 2025. DepthSplat: connecting Gaussian splatting and depth//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 16453-16463 [DOI: 10.1109/CVPR52734.2025.01534]

Xu H, Zhang J, Cai J, Rezatofighi H, Yu F, Tao D, et al. 2023. Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11): 13941-13958 [DOI: 10.1109/TPAMI.2023.3298645]

Xu J, Gao S and Shan Y. 2025. FreeSplatter: pose-free Gaussian splatting for sparse-view 3D reconstruction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE: 25442-25452 [DOI: 10.1109/ICCV51701.2025.02360]

Yang L, Kang B, Huang Z, Xu X, Feng J and Zhao H. 2024. Depth Anything: unleashing the power of large-scale unlabeled data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 10371-10381 [DOI: 10.1109/CVPR52733.2024.00987]

Yu A, Ye V, Tancik M and Kanazawa A. 2021. pixelNeRF: neural radiance fields from one or few images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 4576-4585 [DOI: 10.1109/CVPR46437.2021.00455]

Yu J, Liu X, Luo C, Huang J, Zhou R, Liu Y, et al. 2025. Multitask learning 1997 – 2024: part I fundamentals. Harvard Data Science Review, 7(3) [DOI: 10.1162/99608f92.462f48a2]

Yu T, Kumar S, Gupta A, Levine S, Hausman K and Finn C. 2020. Gradient surgery for multi-task learning//Advances in Neural Information Processing Systems, 33. Red Hook: Curran Associates: 5824-5836 https://proceedings.neurips.cc/paper_files/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html

Zeng Z H, Wang Z J, Zhang Y B, Cai W N, Zhang L L, Guo Y, et al. 2025. Semantic and consistent neural radiance field reconstruction method based on intrinsic decomposition via classification. Journal of Image and Graphics, 30(2): 559-574 (曾志鸿,王宗继,张源奔,蔡伟南,张利利,郭岩,等. 2025. 利用本征属性分类的神经辐射场视角及语义一致性重建. 中国图象图形学报, 30(2): 559-574) [DOI: 10.11834/jig.240140]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhao X, Zhang P, Liu B and Wu Y. 2025. DGOcc: depth-aware global query-based network for monocular 3D occupancy prediction. Neurocomputing, 644: 130426 [DOI: 10.1016/j.neucom.2025.130426]

Zheng Y, Jiang Z, He S, Sun Y, Dong J, Zhang H, et al. 2025. NexusGS: sparse view synthesis with epipolar depth priors in 3D Gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 26800-26809 [DOI: 10.1109/CVPR52734.2025.02496]

Zhou T, Tucker R, Flynn J, Fyffe G and Snavely N. 2018. Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics, 37(4): 65:1-65:12 [DOI: 10.1145/3197517.3201323]

Zhou Y, Lei T, Liu H, Du N, Huang Y, Zhao V, et al. 2022. Mixture-

© 中国图象图形学报版权所有

of-experts with expert choice routing//Advances in Neural Information Processing Systems，35. Red Hook：Curran Associates：7103-7114［DOI：10.52202/068431-0515］

作者简介

焦恺谊，男，硕士研究生，主要研究方向为计算机视觉、三维重建。Email：2024200815@buct.edu.cn

© 中国图象图形学报版权所有